1566 La Pradera Dr
Campbell, CA 95008
www.videoclarity.com
408-379-6952

How Many Humans Does it Take to Judge Video Quality?

Bill Reckwerdt, CTO
Video Clarity, Inc.

Abstract for Subjective Video Quality Assessment

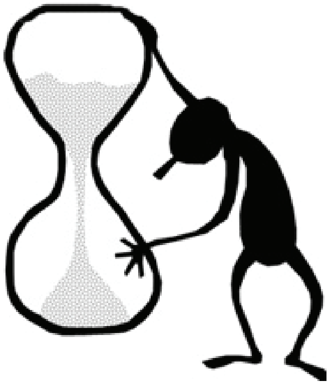
In order to reach your home PC or TV, video sequences go through many stages. They are acquired via a camera, produced/edited, compressed and transmitted. A number of video quality compromises occur at each step. How can you tell if the video quality is good enough? One obvious way is to solicit the opinion of a group of human observers. After all, the video sequences are meant for humans.

What Causes Artifacts?

Camera optics and production/editing affect the video sequence in minimal ways. If care is taking to keep the camera in focus and to produce video sequences which are the same resolution, then these effects are minimized. For this paper, we will assume that a pristine video sequence exits production/editing.

In an ideal world, the pure, uncompressed video sequence is transmitted to the home. To transmit a standard NTSC video, the transmission media needs to handle 720(w)*480(h)*30(fps)*2(bytes) = 21MB/s and that does not include the audio or any ancillary data (like subtitles). The total rate is closer to 27MB/s. To exasperate the problem, HDTV is six times bigger (160MB/s). A good home connection is around 1MB/s (8Mbps). So how do you send 160MB/s down a 1MB/s pipe?

Figure 1: Squeeze Video through the Pipe



To send this data compression must be used. Moreover, the compression must be about 160:1 in the case of High Definition.

The best video quality at a given bit-rate is the goal of video compression. The general idea is to take advantage of the eye and brain. The human eye and brain can fill in details automatically based on experience. Thus, compression techniques take advantage of this innate power. The color space is reduced from millions of colors to a smaller set and brightness is given a higher weight compared to color. Technically, this is why video uses Y’CbCr as opposed to RGB color space. Y’CbCr uses 2-bytes (in 4:2:2) or 12-bits (in 4:2:0) compared to 3-bytes in RGB (4:4:4). Next the video sequences are transformed to the frequency space as groups of pixels forming 8x8, 4x8, etc. sized blocks. Each of these blocks is individually compressed. The blocks are combined together into a macroblocks as a basis for interframe compression - compression between frames (aka temporal compression). At the block boundaries, visible artifacts can form. These artifacts are caused by differences in the luminance or chrominance values at block boundaries. To reduce these artifacts, video sequences are blurred to make the edges softer. Of course, this is just another form of artifact. Remember that all of this is done to trick the eye because the transmission media cannot handle uncompressed video sequences.

On the transmission side, more errors can occur. The Internet can re-order or lose packets randomly. Cloud cover can affect Satellite transmissions, and squirrels can eat away at your underground cables. These errors usually are detected and concealed. The concealment results in frozen video sequences, which won't be observed if the freeze is short enough. If the error is too long, the result will be a mix of frozen images and green blocks.

So How Do You Judge the Video Quality?

Since artifacts can enter the system at various stages, care must be taken to find the cause of the errors. The easy thing to do is to view the video sequence after each stage and judge whether the quality is good. However, this is a very time and resource consuming process. Moreover, what does good mean? A better term is probably is the video quality adequate.

The Video Quality Expert Group (VQEG) created a specification for subjective video quality testing under ITU-R BT.500 recommendation. This recommendation describes methods for subjective video quality analysis where a group of human testers analyze the video sequence and grade the picture quality. The grades are combined, correlated and reported as Mean Opinion Score (MOS).

The main idea can be summarized as follows:

- Choose video test sequences (known as SRC)
- Create a control test environment (known as HRC)
- Choose a test method (DSCQS SSCQE)
- Invite a sufficient number of human testers (20 or more)
- Carry out the testing
- Calculate the average score and scale

Quick Review

Since compression is at least 160:1, the video quality will not be perfect. Comparing the original sequence to the compressed sequence and simply reporting that they are different is not the goal. The goal is to say that the video quality after compression is good enough – not perfect.

Choosing Video Test Sequences

Basically, this means that you must test typical video sequences. If a news channel is being tested for video quality, then the content is mainly faces, graphs and some scenery. On the other hand, a sports channel has quick moving action and crowds of people waving and chanting. So care must be taking to define a set of video sequences that reflects an adequate diversity in video content.

VQEG offers a set of generally hard to compress standard and high-definition video sequences.

Regardless, the reference video sequence should be scaled to the display screen size so that the display graphics are not re-scaling the video sequence. Further, compression should take place on the scaled reference video sequences to create a fair comparison.

Creating a Test Environment

The hardware environment under test is quite likely the video encoder, transmission media, set-top box/PC, and a display. Each display device should be calibrated and if multiple displays are used they must be the same model. The human testers stand a preset distance from the display (usually 3 times the height of the display).

Choose a Test Method

In principle there are two subjective methods for picture quality assessment:

- the DSCQS (Double Stimulus Continual Quality Scale) method
- the SSCQE (Single Stimulus Continual Quality Evaluation) method

The two methods basically differ in that one method makes use of a reference video signal and the other does not. The human testers assess the video sequence (SSCQE) or compare the video sequence to the original video sequence (DSCQS) and issue quality scores from 0 (Bad) to Max (Excellent). Max is 5, 7, 9, or 100 depending on the type of test.

The DSCQS method is preferred when the quality of the test and reference sequences are similar. A small subset of the video sequence can be shown (typically 10 seconds) and small differences in video quality can be easily discerned. Long sequences are harder to evaluate as the human testers become fatigued seeing 2 sets of video sequences for each test.

The SSCQE method has the human testers watching 20-30 minutes of a single video sequence. This method is ideal when measuring the experience of the home user, but it is harder to judge. After all what is good enough? Program content tends to significantly influence the SSCQE scores. Another interesting point is that human memory is not symmetrical. Humans are quick to criticize degradation in video quality; while slow to reward improvements.

Inviting Human Testers and Training

The human testers should be tested for vision problems and briefed about the goals of the experiment. A short training showing a range of video sequence along with video quality scores is used to set target expectations.

Calculate the Average Score and Scale

Although subjective testing conforms to a specification – ITU-R BT.500 Recommendation – successive tests will result in different test scores. The human testers' subjective scale, expectation, and experience will influence the score. Thus, even when the same subjective test methodology is applied, there will be some gain and shift in the scores. A secondary test rating which eliminates data that is outside of the average (inexperience human tester) and linearly scales (higher or lower) using regression techniques must be performed to normalize the data.

How about using an Algorithm to Measure Video Quality Objectively?

Objective Video Quality Measurement seeks to determine the quality of the video sequences algorithmically. While algorithms can track the sum of differences between the reference and compressed video sequences, this is not the goal. The goal is to check if the video quality is good enough after compression. To date, video quality algorithms try to model with good correlation the subjective scores of human testers. This is a little unintuitive. How do you objectively score a subjective test?

Much like Subjective Testing methods defined above, Video Quality Assessment algorithms can be classified into 3 categories:

- Full Reference (FR) methods compare the reference (perfect) video sequence to the processed (distorted) video sequence. Full Reference is generally used as a tool when designing video processing algorithms and when assessing video quality equipment vendors. Examples of Full Reference algorithms include Sarnoff JND, PSNR, SSIM.
- No Reference (NR) methods estimate the distortion level in the video sequence with no knowledge about the original sequence. No Reference algorithms can be used in any settings, but the algorithms are inflexible in making accurate quality predictions.
- Reduced Reference (RR) methods use partial reference information to judge the quality of the distorted signal. If a side channel is present, then Reduced Reference can be used to monitor video quality.

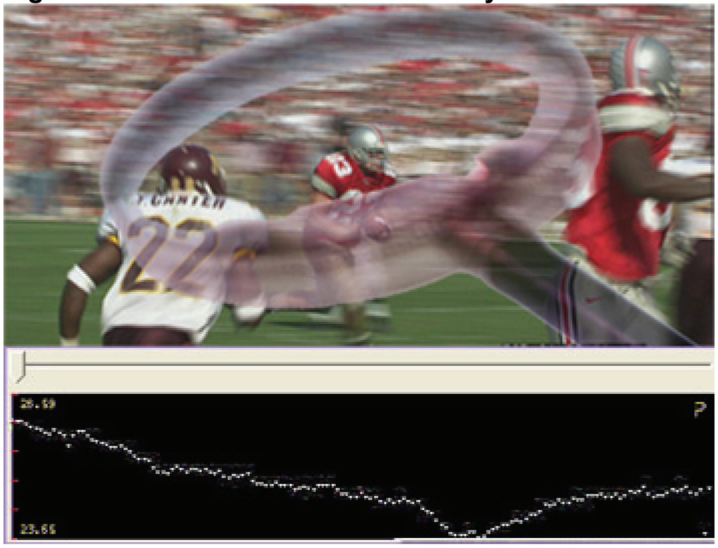
What is Video Clarity?

Video Clarity provides a frame work for video quality testing. Video Clarity's ClearView system combines a video server, video recorder, and file decoder with multiple viewing modes and objective metrics.

ClearView captures video content from virtually any source-file, digital or analog source such as SDI, HD-SDI, DVI, VGA, HDMI, component, composite, or S-video. Regardless of the input, ClearView converts the video sequence to fully uncompressed 4:2:2 Y'CbCr or 4:4:4 RGB. This allows different compression algorithms to be compared and scored relative to each other.

ClearView includes many Objective Metrics to mimic the human visual system. The most famous metrics are Sarnoff's JND and PSNR. The video sequences are normalized, aligned both spatially and temporally, and return an automated, repeatable, and inherently objective pass/fail score, which can be run on any single or series of video frames. Moreover, the scores along with the original video sequences can be shown in multiple viewing modes with VCR-like controls. This allows the operator to see why the objective metric scored the video sequences.

Figure 2: Score the Video and See Why!



Have you ever wanted to compare your H.264/VC-1 & MPEG-2 algorithms relative to each other? Now you can? You can even measure video delay and audio/video lip-sync.

The Author

Bill Reckwerdt has been involved in digital video since the early 90's from digital compression video on demand, to streaming servers. He received his MS specializing in Behavioral Modeling and Design Automation from the University of Illinois Urbana-Champaign. He is the VP of Marketing and the CTO for Video Clarity, which makes quantitative, repeatable video quality testing tools - http://www.videoclarity.com.